DOCUMENT RESUME

ED 073 128 TM 002 372

AUTHOR	Linn, Robert L.
TITLE	Some Implications of the Griggs Decision for Test Makers and Users.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RM-72-13
PUB DATE	11 Nov 72
NOTE	16p.; Paper presented at the Conference on Equal Employment Opportunity and Psychological Testing, New York, November 11, 1972
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Court Litigation; Culture Free Tests; Employment Qualifications; *Equal Opportunities (Jobs); Minority Groups; Negroes; *Personnel Selection; Promotion (Occupational); *Test Bias; *Test Construction; *Test Selection; Test Validity
IDENTIFIERS	*Civil Rights Acts of 1964 Title VII

AESTRACT

The "Griggs vs. Duke Power" decision of the Supreme Court specifies that job relatedness is the only lawful reason for using tests for the purpose of selection when their use results in the disproportionate exclusion of minority group members. One of the main implications of the "Griggs" decision for the test producer and user alike is a renewed emphasis on sound practices of employee selection. The producer of employment tests must put more effort into job analysis activities, criterion development activities, and criterion related validity studies. Another general concern for the test publisher is about certain characteristics of the test or its administration that are irrelevant to the specific abilities being measured but which depress scores for members of one group. Another consideration is that of differential validity studies. There is more than one reasonable definition of test fairness, and these definitions are in conflict; thus, competing values must be weighed. The test maker cannot resolve the problem of competing values, but he can provide the test user with information that will make explicit the value implications of various uses of tests. Two figures illustrate Thorndike's alternate definition of culture fairness. (DB)

ED 073128

RESEARCH MEMORANDUM

RM-72-13

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

TM 002 372

SOME IMPLICATIONS OF THE GRIGGS DECISION FOR TEST MAKERS AND USERS

Robert L. Linn

Paper presented at the Conference on Equal Employment Opportunity and Psychological Testing, sponsored by the Metropolitan Applied Research Center, Inc., Barbizon Plaza Hotel, New York, November 11, 1972.

Educational Testing Service
Princeton, New Jersey
December 1972

Some Implications of the Griggs Decision for Test Makers and Users

Robert L. Linn

In the Griggs vs. Duke Power decision the Supreme Court clearly identified job-relatedness as the only lawful reason for using tests for the purpose of selection when their use results in the disproportionate exclusion of minority group members. In the unanimous decision delivered by Chief Justice Burger, he wrote: "If an employment practice which operates to exclude Negroes cannot be shown to be related to job performance, the practice is prohibited."[1] Where tests are demonstrated to be job related, however, the Griggs decision makes it clear that tests are judged to be legal and useful. In the words of Chief Justice Burger, "Nothing in the [Civil Rights] Act [of 1964, Title VII] precludes the use of testing or other measuring procedures; obviously they are useful .... What Congress has commanded is that any tests used must measure the person for the job and not the person in the abstract."[2]

The necessity of demonstrating job-relatedness is clear, but the evidence that is required for the demonstration is less clear. The EEOC Guidelines on Employment Selection Procedures[3] and some of the litigation following the Griggs decision help identify the type of evidence that will probably be required. The EEOC Guidelines are given specific support in the Griggs decision where it is argued that there is "...good reason to treat the Guidelines as expressing the will of Congress."[4] Given this support, the Guidelines would seem to be the natural place for the employer, the test producer, and the personnel psychologist to turn in order to insure that the use of tests for employee selection or promotion is consistent with the requirements of the Civil Rights Act of 1964.

The Guidelines stress the importance of careful job analysis. The job analysis provides the basis for test selection. It also may provide the basis for defending the rational validity of the test or, preferably, the development of appropriate criterion measures to be used in the empirical investigation of criterion-related validity. Bill Enneis has pointed out that the Guidelines "...embody the substance of good personnel employment practices as recommended by experts in industrial psychology and personnel administration for the past forty or fifty years."[5] I'm in general agreement with this evaluation. The reader of the Guidelines can certainly find much in common with professionally accepted prescriptions for personnel selection. While generally agreeing with this position, Guion has identified two aspects of the Guidelines that seem to be new but not "...especially heretical to orthodox testers."[6] The two new features are (1) an expanded definition of tests and (2) the emphasis on independent validation for minority groups (i.e., the emphasis on differential validity studies). I'll have more to say about the emphasis on differential validity a bit later.

From the above perspective, one of the main implications of the Griggs decision for the test producer and the user alike is a renewed emphasis on sound practices of employee selection. It is no longer sufficient to use a test because it reliably measures general ability. It must be shown that the test is measuring an attribute that is important for successful job performance. In some cases professionally developed tests of performance (e.g., a typing test) may stand on the rational relationship of the performance to the analysis of the job. For more general ability tests, however, criterion-related validity or a strong combination of construct validity coupled with an analysis of the importance of the constructs to job performance will probably be required.

To provide the necessary backup for test users the producer of employment tests is going to have to put more effort into job analysis activities, criterion development activities, and criterion related validity studies than many of them have done in the past. For many, if not most jobs, however, the magnitude of the effort that is required is tremendous. Judging from the ETS-Civil Service Commission study of three occupations that was directed by Joel Campbell,[7] it seems unreasonable to expect that the typical local situation will permit the kind of comprehensive study that might be desired. Careful attention to comprehensive job analyses, which were used to develop objective criterion measures and select tests that measured abilities that were judged to be important for good performance on the job, paid off in impressive validities in Campbell's study. However, that study required several years to accomplish as well as substantial support from the Ford Foundation, the close cooperation of the Civil Service Commission, and contributions from a number of ETS and Civil Service staff members. Reflection on the magnitude of the effort in Campbell's study in contrast to the resources that are typically available in the real-life personnel situation led Anastasi to ask: "In a more nearly typical personnel situation, what...can be done to ensure that selection tests are truly valid, or relevant to the job?"[8] In response to her own question Anastasi answered: "For this purpose...I would turn to a thorough, professional job analysis, followed by a study of the published findings regarding the validity of different tests against specific job functions."[9] She urged "... that more effort be expended on basic research regarding the specific aspects of behavior measured by different instruments and less on inadequate, inconclusive, local validation studies against global criteria of job performance."[10] I

would heartily endorse this position. It seems to me to be very consistent with both the need for better construct validity and Lawshe's notion of synthetic validity.[11] It also suggests an important role for the test producer, namely, that of providing the necessary research base required to understand the aspects of behavior that are measured by their instruments so that the personnel psychologist will have a sound basis for using his job analysis to select tests.

While obviously stimulated by the need to ensure that employment tests are used appropriately with minority groups, the above concerns and recommendations are actually quite general and should apply to all testing--not just testing of minority group members. Another general concern for the test publisher that is given salience by questions of equal employment opportunities for minority group members is the concern about certain characteristics of the test or its administration that are irrelevant to the specific abilities being measured but which depress scores for members of one group. The classic example is the test that requires considerable verbal ability or has difficult reading requirements when these are not the skills that are being measured. The context of the test items or specific words may be more familiar to one group than another, and where these factors are not essential to the attribute being measured, they should be eliminated. One way to ensure this is for the test producers to obtain the input from representatives of the various groups who will be taking the test. In other words, minority group members need to be represented in all phases of the test construction process (e.g., test specifications, item writing, and reviews of the test). Comparative item analyses for different groups of people are also potentially useful for this purpose.

Another consideration is that of differential validity studies. As was mentioned earlier, the EEOC Guidelines stress the importance of such studies. "Data must be generated and results separately reported for minority and nonminority groups wherever technically possible."[12] I would not argue against the desirability of differential validity studies where feasible; however, it may be worthwhile to consider the implication of evidence that is accumulating from the differential validity studies that have been conducted to date. Obviously the evidence is not complete, but a fairly substantial number of studies have been conducted in the past few years.

At the recent meetings of the American Psychological Association, William Ruch reviewed differential validity studies that were conducted in a nonmilitary business or industrial situation and that had separate statistics reported for blacks and whites which permitted him to test for homogeneity of regression for the two groups. He also required that race not be confounded with some other variable (e.g., blacks working for one company and whites for another). From his analyses, Ruch concluded: "Certainly these 20 studies do not tell the whole story. The evidence they do provide is that there is no such thing as differential validity but there is a tendency of tests to overestimate black job performance."[13]

Ruch is by no means the only person to conclude that tests which are valid for one group are usually valid for the other (i.e., differential validity is rare) and that where differences in the prediction systems are found, the predictions based on the total group usually tend to overestimate rather than underestimate minority group performance on the criterion. For example, this conclusion is reached by Bray and Moses in their chapter in the 1972 Annual Review of Psychology,[14] and is one of the clear results of Campbell's Civil Service Study[15] which I mentioned

previously. In a military setting, a similar conclusion is reported by Guinn, Tupes and Alley,[16] and in an educational setting the tendency is also found to be in the direction of overprediction.[17]

Inasmuch as the overprediction finding is valid, it has implications that could run counter to one of the intents of the Civil Rights Act and to the reason for stressing differential validity studies in the Guidelines. If cutoff scores are "...set so as to predict the same probability of success in both groups"[18] as the Guidelines suggest, and the majority group regression line overpredicts for the minority group, the cutoff score would be set higher for the minority group than for the majority group. Thus, the differential validity concept which was intended to have a favorable influence on the employment of minority group members could be used to defend a policy that actually excluded more minority group members than would have been the case previously.

There are many reasons why the above use of such differential validity study results should be emphatically rejected. Werts and I have argued elsewhere that there are statistical and psychometric reasons that could lead one to expect the overprediction result.[19] These are: (1) the failure to include a variable in the prediction system on which the groups differ and (2) the lack of perfect test reliability. There are also less subtle artifacts that could cause the overprediction result, for example, the criterion variable itself may not be free of bias. This is a particularly plausible explanation when supervisor ratings are used as the criterion variable.

In my view, however, there is a still more important reason for rejecting the possible implication of differential validity results that higher cutoff scores be required of the minority group than the majority

group. This reason is that the maximization of performance on the job is only one value that needs to be attended to, and perhaps not the most important one.

The usual definition of test fairness in terms of equal regression equations is fair to the institution in the sense that it provides the institution with a way of selecting people such that the average criterion score of those selected is a maximum. It is also "fair" to individual members of the two groups in that the criterion performance is not systematically over- or underpredicted for members of either group. Nonetheless, Thorndike clearly demonstrated that a test which has equal regression equations for the two groups "...is 'unfair' to the lower scoring group as a whole in the sense that the proportion qualified on the test will be smaller, relative to the higher-scoring group, than the proportion that will reach any specified level of criterion performance."[20]

Thorndike has argued persuasively that the traditional way of looking at test fairness in terms of equality of the prediction systems is not the only perspective that should be considered. The problem with the traditional definition that Thorndike has so clearly identified is illustrated in Figure 1. In Figure 1 a situation is depicted in which the regression of the criterion on the test is identical in the two groups (shown by the solid line with a slope equal to .25). Thus, the test would be considered "fair" from the traditional psychometric point of view and according to the Guidelines. Note, however, that the difference in means on the test is substantially larger than the difference on the criterion variable.

Insert Figure 1 about here

Now suppose for purposes of illustration that the mean of group A on the criterion variable corresponded to what was considered minimum acceptable performance (i.e., those with criterion scores above that point are considered successes and those below it are considered failures). As can be seen in Figure 1, about 50% of the hypothetical group A have criterion scores above this cutting point and about 20% of group B (shaded area of the group B distribution curve) have criterion scores above the cutting point. Since criterion scores are not known in advance, it is the predicted criterion scores that are used for selection. If only those individuals with predicted scores equal to the success-failure point on the criterion variable were accepted, then approximately 50% of group A would be accepted but essentially none of the group B members would be accepted. This is the phenomenon that Thorndike was referring to when he said that while the traditional approach with equal regression equations is "fair" to individual members of the minor group, it is "unfair" to the minor group as a whole, "...in the sense that the proportion qualified on the test, relative to the higher scoring group, will be smaller than the proportion that will reach any specified level of criterion performance."[21] In the illustration depicted in Figure 1, the relative proportions in groups A and B that are qualified on the test are .50 and essentially zero, respectively, whereas the proportions above the success-failure point on the criterion variable are approximately .50 and .20 respectively.

Thorndike proposes that: "An alternate definition [of fairness] would specify that the qualifying scores on the tests should be set at levels that will qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified level of criterion performance."[22] Thus, we have two conflicting definitions of fairness. The

only time that the equal regression definition and Thorndike's definition are in agreement is when the validity is perfect. This unrealistic case is depicted by the 45° line in Figure 2. With perfect validity and the same regression equation in both groups, the relative proportions qualified on the test (i.e., above the accept-reject point in Figure 2) are equal to the relative proportions that are successful (i.e., above the success-failure point in Figure 2).

Insert Figure 2 about here

The requirement that an equal proportion of each group be accepted as would be successful also could be satisfied by a situation where the regression line for the group B lies below the one for group A (i.e., the group A equation overpredicts for group B) but the cutting score appropriate for group A is used to select within both groups. This situation is also depicted in Figure 2 by the two parallel lines with slopes of .25.

In summary, Thorndike has identified a flaw in the traditional definition of "fairness." His formulation suggests that we should be looking at the implications for the proportions of students admitted as well as the regression lines. For a case with the same regression equations as that illustrated in Figure 1, different cutting scores would be required to make the proportions qualified on the test equal to the proportions successful on the criterion. In particular, a lower cutting score on the test would be required for the hypothetical group B in Figure 1 in order to satisfy Thorndike's alternate definition.

The problem that Thorndike identified is implicit in some earlier discussions. For example, Cooper and Sobel seem to be referring to this problem in their discussion of tests with low, but significant, validity. They argue that where such tests have an "...adverse impact on blacks, use of test scores as a major factor in employment decisions is likely to be unnecessarily prejudicial to blacks."[23] More than one sense of fairness seems implicit in Cooper and Sobel's comment.

In my opinion, Thorndike's main contribution is that of making it explicit that there is more than one reasonable definition of test fairness and that these definitions are in conflict. Thus, we must look beyond a simple technical resolution of the problem. Competing values must be weighed. Errors of selecting individuals who are unsuccessful and errors of rejecting applicants who would be successful must be weighed for members of minority groups and for members of the majority group. The former error is of prime concern in some employee selection situations where mistakes on the job can be very costly or even disastrous (e.g., the selection of pilots). In many situations, however, the latter type of error may be of greater concern. The test maker cannot resolve the problem of competing values, but he can provide the test user with information that will make explicit the value implications of various uses of tests.

Footnotes

[1]United States Supreme Court, Griggs vs. Duke Power, March 8, 1971.

[2]Ibid.

[3]Equal Employment Opportunity Commission. Part 1607-Guidelines on Employee Selection Procedures. Code of Federal Regulations, Title 29, Chapter XIV, 1970.

[4]United States Supreme Court, op.cit.

[5]Enneis, W. H. Statement before the House Post Office and Civil Service Subcommittee. Reprinted in B. R. Anderson and M. P. Rogers (Eds.), Personnel Testing and Equal Employment Opportunity. Washington, D. C.: U. S. Government Printing Office, 1970. Pp. 16-20.

[6]Guion, R. M. Recruiting selection and job placement. In M. D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology, Chicago, Illinois: Rand McNally (in press).

[7]Campbell, J. T. Principal results of study and conclusions. In Lois Crooks (Ed.), An investigation of sources of bias in the prediction of job performance....a six year study. Proceedings of Invitational Conference held June 22, 1972, New York. Princeton, New Jersey: Educational Testing Service, 1972. Pp. 9-77.

[8]Anastasi, Anne. Sources of bias in the prediction of job performance: Technical critique. In Lois Crooks (Ed.), An investigation of sources of bias in the prediction of job performance....a six year study. Proceedings of Invitational Conference held June 22, 1972, New York. Princeton, New Jersey: Educational Testing Service, 1972. Pp. 79-88 (p. 87).

[9]Ibid. Pp. 87-88.

[10]Ibid. P. 88.

[11]Lawshe, C. H. Employee Selection. Personnel Psychology, 1952, 5. Pp. 31-34.

[12]Equal Employment Opportunity Commission, op.cit.

[13]Ruch, W. W. A re-analysis of published differential validity studies. Paper presented at the Symposium on: Differential Validation under EEOC and OFCC Testing and Selection Regulations, American Psychological Association, Honolulu, Hawaii, September 6, 1972.

[14]Bray, D. W. & Moses, J. L. Personnel selection. Annual Review of Psychology, 1972, 23. Pp. 545-576.

[15]Campbell, J. T., op.cit.

[16]Guinn, N., Tupes, E. C. & Alley, W. E. Cultural subgroup differences in the relationships between Air Force aptitude composites and training criteria. (Technical Report 70-35.) Lackland Air Force Base, Texas: Human Resources Research Center, September 1970.

[17]Cramer, S. & Stevic, R. R. A review of the 1971-72 literature: Research on the transition from high school to college. College Board Review, 1972, 85. Pp. 32-38.

[18]Equal Employment Opportunity Commission, op.cit.

[19]Linn, R. L. & Werts, C. E. Considerations for studies of test bias. Journal of Educational Measurement, 1971, 8. Pp. 1-4.

[20]Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8. Pp. 63-70 (p. 63).

[21]Ibid. P. 63.

[22]Ibid. P. 63.

[23]Cooper, G. & Sobel, R. B. Seniority and testing under fair employment laws: A general approach to objective criteria. Harvard Law Review, 1969, 82. Pp.1598-1679 (p. 1663).

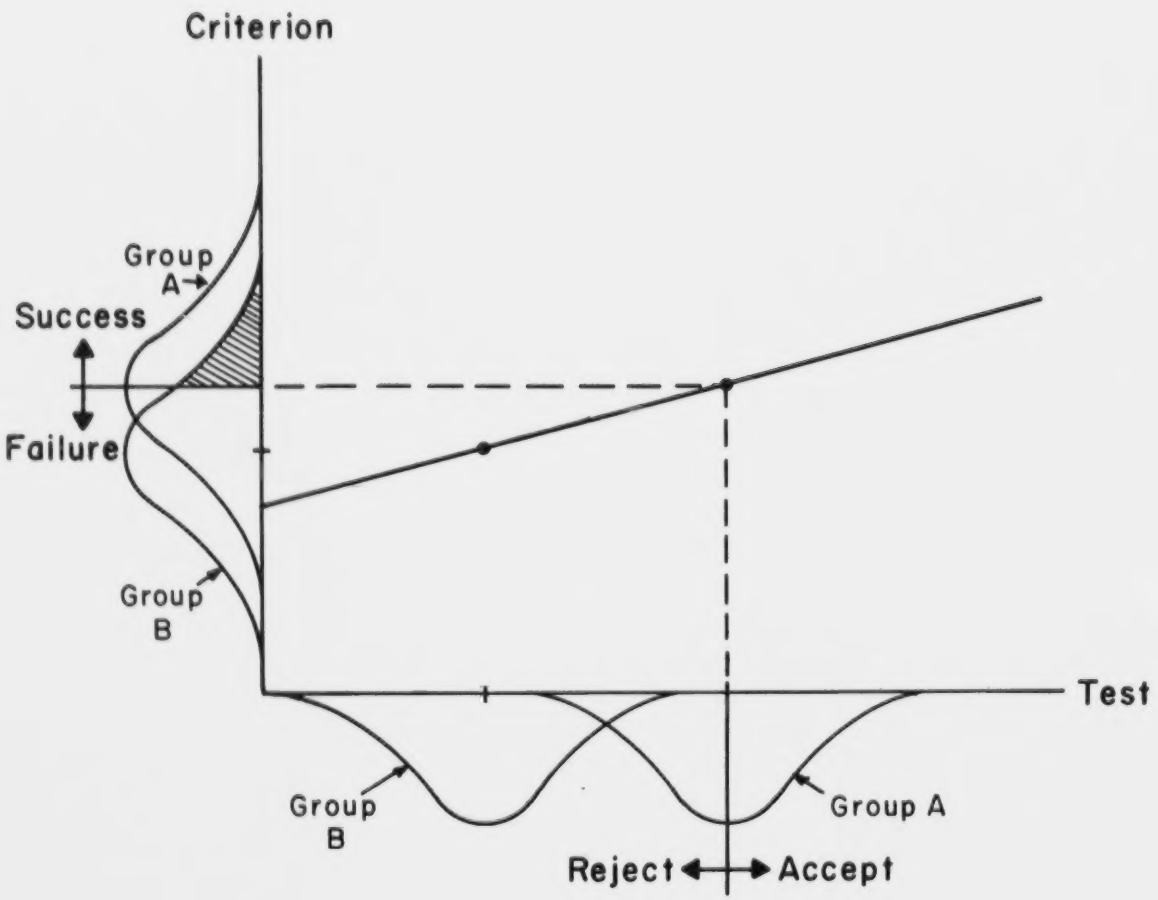

Figure I
Illustration of Thorndike's Alternate Definition
of Culture Fairness

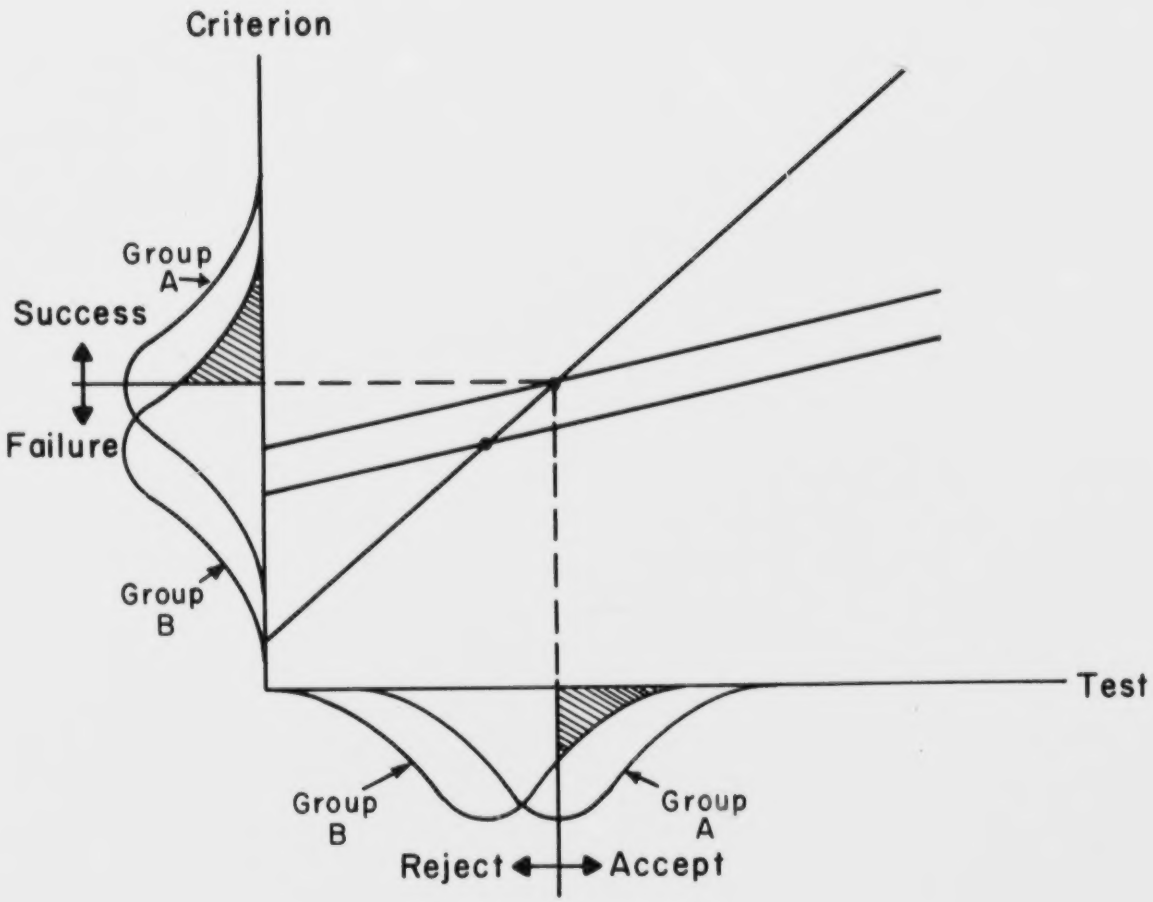

Figure 2

Illustration of Thorndike's Alternate Definition of Culture Fairness